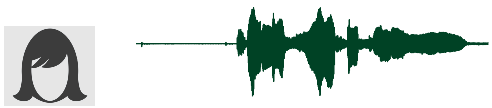
SD-QA: Spoken Dialectal Question Answering for the Real World

Fahim Faisal, Sharlina Keshava, Md Mahfuz ibn Alam, Antonios Anastasopoulos
Department of Computer Science, George Mason University
`{ffaisal,skeshav,malam21,antonis}@gmu.edu`

Abstract

Question answering (QA) systems are now available through numerous commercial applications for a wide variety of domains, serving millions of users that interact with them via speech interfaces. However, current benchmarks in QA research do not account for the errors that speech recognition models might introduce, nor do they consider the language variations (dialects) of the users. To address this gap, we augment an existing QA dataset to construct a *multi-dialect, spoken* QA benchmark on four languages (Arabic, Bengali, English, Kiswahili) with more than 68k audio prompts in 22 dialects from 245 speakers. We provide baseline results showcasing the real-world performance of QA systems and analyze the effect of language variety and other sensitive speaker attributes on downstream performance. Last, we study the fairness of the ASR and QA models with respect to the underlying user populations.[1]



↓ ASR

Transcription: `Teolojia ya dogma ni nini?`

↓ QA

Minimal Answer: `neno linalotumika hasa kumaanisha fundisho la imani lisiloweza kukanushwa na wafuasi wa dini fulani.`

Figure 1: Illustration of the envisioned scenario for a user-facing QA system that SD-QA aims to evaluate (example from Swahili).

1 Introduction

The development of question answering (QA) systems that can answer human prompts with, or in some cases without, context is one of the great success stories of modern natural language processing (NLP) and a rapidly expanding research area. Usage of such systems has reached a global scale in that millions of users can conveniently query voice assistance like Google Assistant,[2] Amazon Alexa,[3] or Apple Siri[4] through their smartphones or smart home devices. QA systems have also made large strides in technology-adjacent industries like healthcare, privacy, and e-commerce platforms.

Wide-spread adoption of these systems requires they perform consistently in real-world conditions, an area where Ravichander et al. (2021) note there is substantial room for improvement. Existing QA system evaluation benchmarks rely on text-based benchmark data that is provided in written format without error or noise. However, inputs to real-world QA systems are gathered from users through *error-prone interfaces* such as keyboards, speech recognition systems that convert verbal queries to text, and machine-translation systems. Evaluating production-ready QA systems on data that is not representative of real-world inputs is problematic and has consequences on the utility of such systems. For example, Ravichander et al. (2021) quantify and illustrate the effects of interface noise on English QA systems.

In this work, we address the need for realistic evaluations of QA systems by creating a multilingual and multi-dialect *spoken* QA evaluation benchmark. Our focus is the utility of QA systems on users with varying demographic traits like age, gender, and dialect spoken. Our contributions are as follows:

1. We augment the TyDi-QA dataset (Clark

[1] The dataset and code for reproducing our experiments are available here: https://github.com/ffaisal93/SD-QA.

[2] https://assistant.google.com/

[3] https://developer.amazon.com/alexa

[4] https://www.apple.com/siri/

et al., 2020) with spoken utterances matching the questions. In particular, we collect utterances in four languages (Arabic, Bengali, English, Kiswahili) and from multiple varieties[5] (seven for Arabic, eleven for English, and two for each of Bengali and Kiswahili).

2. We perform contrastive evaluation for a baseline pipeline approach that first transcribes the utterances with an ASR system and then provides the transcription to the QA system. We compare general and localized ASR systems, finding wide divergences between the downstream QA performance for different language varieties.

2 The SD-QA dataset

The SD-QA dataset builds on the TyDi-QA dataset (Clark et al., 2020), using questions, contexts, and answers from four typologically diverse languages. Our focus is to augment the dataset along two additional dimensions. The first is a speech component, to match the real-world scenario of users querying virtual assistants for information. Second, we add a dimension on dialectal and geographical language variation.

2.1 Languages and Varieties

We focus on four of the languages present in the original TyDi QA dataset: English, Arabic, Bengali, and Kiswahili.[6] In total, SD-QA includes more than 68k audio prompts in 22 dialects from 245 annotators. Table 1 presents a list of the different locations from where we collected spoken samples. A detailed breakdown of dataset and speaker statistics is available in Appendix A. We provide a brief discussion on the dialectal variation exhibited in the dataset's languages in the following paragraphs. We note that English and Arabic varieties are over-represented, but this is due to cost: it was easier and cheaper to source data in English and Arabic than in other languages; we plan to further expand the dataset in the future with more varieties for Bengali and Kiswahili as well as more languages.

Arabic is practically a family of several varieties, forming a dialect continuum from Morocco to the west to Oman in the east. Most of NLP has focused on Modern Standard Arabic (MSA), the literary standard for language of culture, media, and education. However, no Arabic speaker is really a MSA native speaker– rather, most Arabic speakers in every day life use their native dialect, and often resort to code-switching between their variety and MSA in unscripted situations (Abu-Melhim, 1991). We follow a sub-regional classification for the Arabic dialects we work with, as these dialects can differ in terms of their phonology, morphology, orthography, syntax, and lexicon, even between cities in the same country. We direct the reader to (Habash, 2010) for further details on Arabic handling in NLP, and to (Bouamor et al., 2018) for a larger corpus of Arabic dialects.

The original TyDi-QA dataset provides data in MSA, as the underlying Wikipedia data are in MSA.[7] To capture the proper regional variation of Arabic as much as possible, the instructions to the annotators were to first read the question and then record the same question using their local variety, rephrasing as they deem appropriate.

Bengali is an Indo-Aryan language spoken in the general Bengal region, with the majority of speakers concentrated in modern-day Bangladesh and the West Bengal state in India (Klaiman, 1987). The language exhibits large variation,

Language	Locations (Variety Code)
Arabic	Algeria (DZA), Bahrain (BHR), Egypt (EGY), Jordan (JOR), Morocco (MAR), Saudi Arabia (SAU), Tunisia (TUN)
Bengali	Bangladesh-Dhaka (BGD), India-Kolkata (IND)
English	Australia (AUS), India-South (IND-S), India-North (IND-N), Ireland (IRL), Kenya (KEN), New Zealand (NZL), Nigeria (NGA), Philippines (PHI), Scotland (SCO), South Africa (ZAF), US-Southeast (USA-SE)
Kiswahili	Kenya (KEN), Tanzania (TZA)

Table 1: Languages and sample collection locations (roughly corresponding to different spoken varieties) in the SD-QA dataset.

[5]We will use the terms "dialect" and "language variety" interchangeably.

[6]The languages were selected for their typological diversity and the wide range of variation they exhibit.

[7]While Clark et al. (2020) do not specify whether the data are in MSA or in any vernaculars, we rely on a native speaker for this observation.

Dhaka	Kolkata	English
দোয়া *(doa)*	প্রার্থনা *(prarthona)*	pray
পানি *(pani)*	জল *(jol)*	water
দাওয়াত *(daowat)*	নিমন্ত্রণ *(nimontron)*	invitation

Table 2: Example of lexical differences between the Dhaka and Kolkata Bengali varieties.

although languages like Sylheti and Chittagonian, once considered varieties of Bengali, are now considered as languages on their own (Simard et al., 2020; Masica, 1993). We focus our collection on two major geographical poles: Dhaka (Bangladesh) and Kolkata (India). The Dhaka variety is the most widely spoken one, while the Rarhi variety (Central Standard Bengali) is prevalent in Kolkata and has been the basis for standard Bengali. Thompson (2020) notes that the linguistic differences between the two sides of the border are minor. The differences are mainly phonological, although some lexical differences are also observed in common words. For example, the bangla spoken in Dhaka dialect is more spread out like using চইলা *(choila, to go)*, কইরা *(koira, to do)* instead of using চলে *(chole, to go)*, করে *(kore, to do)*Arabic influence is prevalent in Dhaka dialect, whereas the Kolkata dialect is more influenced by other Indo-Aryan languages, resulting in lexical differences (examples shown in Table 2).

English is a West Germanic Indo-European language with a very wide geographic distribution. This is attributed to colonialism, and resulted in large differences in grammatical patterns, vocabulary, and pronunciation between English varieties. For our corpus, it was only feasible to sample from a subsection of the English varieties that exist globally. We include regions where English is an “official" language (meaning it is generally the language of the government and of instruction in higher education): Australia, India, Ireland, Kenya, New Zealand, Philippines, Scotland, South Africa, and the United States. We note that there are important differences between English usage in these regions. For example, even though English is an official language in India and Kenya, speakers are more likely to use it as a second language, while having a different native language.

Kiswahili (or kiSwahili or Swahili, ISO code: swa) is a Bantu language that functions as a *lingua franca* for a large region of central Africa, as it is spoken in Tanzania, Kenya, Congo, and Uganda. Its grammar is characteristically Bantu, but it also has strong Arabic influences and uses a significant amount of Arabic loanwords. While there are more than a dozen Swahili varieties, the three most prominent are kiUnguja, spoken on Zanzibar and in the mainland areas of Tanzania, which is also the basis of considered-standard Swahili; kiMvita, spoken in Mombasa and other areas of Kenya; and kiAmu (or Kiamu), spoken on the island of Lamu and adjoining parts of the coast. Also prominent is Congolese Swahili (ISO code: swc), which we treat as a separate language because of its significant French influences (due to colonisation). In this work we collected utterances from Kenya (Nairobi area) and from Tanzania (Dar-es-Salaam area). Some of our contributors self-reported as non-native Swahili speakers, naming languages like Kikuyu (ISO code: kik) and Chidigo (ISO code: dig) as their native ones.

2.2 Data Collection Process

Data was collected through subcontractors. Each annotator was a native speaker of the language who grew up and lived in the same region we were focusing on. The annotators were paid a minimum of $15 per hour.[8] We aimed for gender- and age-balanced collection. The data for almost all dialects are gender-balanced,[9] but not all age groups are represented in all dialects (e.g. all Kenyan English and Jordan Arabic speakers are in the 18–30 age group, while all Swahili speakers and Scottish English speakers are in the 31–45 group).

For the collection of the Bengali and Swahili data we used the LIG-Aikuma mobile application (Gauthier et al., 2016) under the elicitation mode. The annotators were shown one question at a time, and they were instructed to first read the question in silence, and then read it out loud in a manner similar to how they would ask a friend or query a virtual assistant like Google Assistant, Siri, or Amazon Alexa.

Data selection and Splits We perform data selection and partitioning by following the process detailed in XOR-QA (Asai et al., 2020), another TyDi-QA derivative dataset. We use the

[8] We note, though, that no annotator needed more than 40 minutes for recording the maximum of 300 questions that corresponded to them.

[9] We note than none of our annotators self-reported as non-binary or other gender beyond male or female.

development set of the original TyDi-QA as our test set, and randomly sample a part of the original TyDi-QA training set for our development set. The development and test partitions for the XOR-QA and our SD-QA dataset are exactly the same for Arabic and Bengali.[10]

As our custom development set is constructed from the TyDi-QA training dataset, the SD-QA training set is constructed by discarding our development instances from the original TyDi-QA training data. We note, though, that we do not provide any spoken QA data for training.

2.3 Limitations

We believe that SD-QA certainly presents an improvement over existing benchmarks with regards to the accurate representation of the underlying users of QA systems in real-world scenarios. However, it is not without limitations. Since we lack access to *actual* spoken queries to smart assistants, we had to resort to recording readings of text questions. Read speech has different characteristics than natural (spontaneous) speech, especially in terms of rate and prosody (Batliner et al., 1995). In the future, we plan to investigate the creation of a spoken QA benchmark following a process similar to the Natural Questions dataset (Kwiatkowski et al., 2019) in order to produce an even more realistic benchmark.

3 Tasks and Evaluation

We perform three tasks over our dataset, defined below. The passage selection and minimal answer selection tasks are directly modeled after the primary task in TyDi-QA:

1. **Automatic Speech Recognition (ASR) Task:** A standard task defined over the utterances of the different language varieties. Given the audio file of the utterance, the model has to produce an accurate transcription. Since our questions are parallel across all language varieties, SD-QA can be used for contrastive evaluation of the robustness of ASR systems across different varieties.

2. **Passage Selection Task:** Given the question and a number of candidate passages, this task asks the model to return the index of the passage containing the answer, or null if no such passage exists.

3. **Minimal Answer Selection Task:** Given the question and a single passage, the task is to return the start and end byte indices of the minimal span that completely answer the question, or a YES/NO answer if appropriate, or NULL if such answer does not exist.

[10]English and Swahili are not part of XOR-QA.

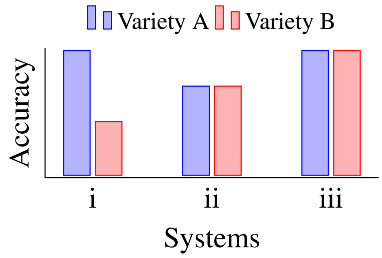


Figure 2: Schematic visualization of three systems tested on two language varieties. System (ii) is preferable to system (i) since it is more equitable. System (iii) is the ideal scenario.

Evaluation The two tasks that are similar to the TyDi-QA primary tasks use the same metric, F_1-SCORE.[11] For the ASR task, we evaluate the models computing the average word error rate (WER) across all test utterances. Unlike TyDi-QA, we (a) include English in our metrics, and (b) will not compute macro-averaged results across all languages as this measure could be biased towards English whose varieties are over-represented, but rather we only report macro-averages across the varieties of each single language.

For the ASR task, we evaluate the models computing the average word error rate (WER) across all test utterances. Unlike TyDi-QA, we (a) include English in our metrics, and (b) only report macro-averages across the varieties of each single language. We elect not to compute macro-averaged results across all languages as this measure could be biased towards English whose varieties are over-represented.

In addition, for all tasks, we will compare the models' robustness and equitability across the varieties. A truly robust, equitable system would perform equally well for all language varieties that it is meant to be used for. Since our dataset is parallel across varieties, it is ideal for *constrastive* evaluations with meaningful comparisons. Conceptually, consider the three

[11]We direct the reader to (Clark et al., 2020) for details.

systems depicted in Figure 2 tested on two language varieties A and B. Systems (i) and (ii) have the same macro-averaged accuracy, but system (ii) is more equitable than system (i): the quality of system (i) in language variety B is worse than in variety A. A more equitable situation is scenario (ii) were the two varieties have about the same performance, but this is achieved at the expense of performance in variety A. Scenario (iii) is an ideal case, where the system can properly model all language varieties and is also equitable. We measure equity by reporting the difference between the minimum and maximum performance attained across a language's variety, as well as standard deviations (in both cases lower is better).

Comparing the Systems' Unfairness When evaluating multilingual and multi-dialect systems, it is crucial that the evaluation takes into account principles of fairness, as outlined in economics and social choice theory (Choudhury and Deshpande, 2021). We follow the least difference principle proposed by Rawls (1999), whose egalitarian approach proposes to narrow the gap between unequal accuracies.

We will measure a system's *unfairness* with respect to the different subgroups using the adaptation of generalized entropy index described by Speicher et al. (2018), which considers equities within and between subgroups in evaluating the overall unfairness of an algorithm on a population. The generalized entropy index for a population of n individuals receiving benefits $b_1, b_2, \ldots, b_n$ with mean benefit μ is

$$\mathcal{E}^\alpha(b_1, b_2, \ldots, b_n) = \frac{1}{n\alpha(\alpha-1)} \sum_{i=1}^{n} \left[\left(\frac{b_i}{\mu} \right)^\alpha - 1 \right].$$

Using $\alpha = 2$ (which we do, following Speicher et al. (2018)), the generalized entropy index corresponds to half the squared coefficient of variation.[12]

If the underlying population can be split into $|G|$ disjoint subgroups across some attribute (e.g. gender, age, or language variety) we can decompose the total unfairness into individual- and group-level unfairness. Each subgroup $g \in G$ will correspond to n_g individuals with corresponding benefit vector $\mathbf{b}^g = (b_1^g, b_2^g, \ldots, b_{n_g}^g)$ and mean benefit μ_g. Then, generalized entropy can be re-written as:

[12]The coefficient of variation is simply the ratio of the standard deviation σ to the mean μ of a distribution.

$$\begin{aligned}\mathcal{E}^\alpha(b_1, b_2, \ldots, b_n) = &\sum_{g=1}^{|G|} \frac{n_g}{n} \left(\frac{\mu_g}{\mu} \right)^\alpha \mathcal{E}^\alpha(\mathbf{b}^g) \\ &+ \sum_{g=1}^{|G|} \frac{n_g}{n\alpha(\alpha-1)} \left[\left(\frac{\mu_g}{\mu} \right)^\alpha - 1 \right] \\ =& \mathcal{E}_\omega^\alpha(\mathbf{b}) + \mathcal{E}_\beta^\alpha(\mathbf{b}).\end{aligned}$$

The first term $\mathcal{E}_\omega^\alpha(\mathbf{b})$ corresponds to the weighted unfairness score that is observed *within* each subgroup, while the second term $\mathcal{E}_\beta^\alpha(\mathbf{b})$ corresponds to the unfairness score *across* different subgroups. This formulation allows us to also study the tradeoff between individual and group-level (un)fairness.

In this measure of unfairness, we define the benefit as being directly proportional to the system's accuracy. For the speech recognition task, we will simply use $b = 1 - WER$ as the benefit that a user receives from that particular interaction with the system. If the system produces a perfect transcription (WER=0) then the user will receive the highest benefit of 1. If the system fails to produce any correct words in the transcription (WER=1) then the user receives no benefit ($b = 0$) from the interaction with the system. For the QA tasks, we make the assumption that the benefit that the user receives is directly proportional to the quality of the answer as measured by F_1-SCORE, so we will use $b = F_1$-SCORE for each question/answer pair.

As we will show in the results section that follows, these unfairness scores are very useful for comparing two systems that can be applied over diverse populations tagged with sensitive features.

4 Baseline Results and Discussion

Baseline Models We benchmark the speech recognition systems using the ASR models through the Google speech-to-text API.[13] For all language varieties, we follow a *pipeline* approach, where we first transcribe the audio utterance and use the output as the input (question) to the QA model. We leave end-to-end multimodal approaches that operate directly on the audio for future work.

Our QA model follows the recipe of (Alberti et al., 2019), training a single model for all languages using multilingual BERT (Devlin et al.,

[13]https://cloud.google.com/speech-to-text

2019) as a feature extractor. The model is trained for a maximum of 10 epochs, selecting the best checkpoint based on development set F_1-SCORE.

4.1 Speech Recognition Results

Transcription Quality In this section, we discuss the quality of the outputs produced by different Automatic Speech Recognition (ASR) units. For English, we transcribe the regional variety utterances using both the localized[14] ASR model e.g. using the `en-AU` system for Australian English and the `en-NZ` for the data from New Zealand. In addition, we also transcribe all English data with the `en-US` system, which will allow us to compare the effectiveness of using localized models versus using a single "general" model. For Arabic we only use the localized model for each variety (as there is no "general" Arabic model available). For Kiswahili and Bengali we use localized models corresponding to the two data collection locations (Kenya-Tanzania and Bangladesh-India respectively) on both collections.

Table 3 presents the WER of the ASR models on the development set for all language varieties in SD-QA. The first important observation is that the average quality between languages varies significantly. The ASR systems achieve the lowest average WER on English (around 11-12), followed by Bengali (~27.6). Arabic and Kiswahili still prove challenging, with WER around 36 and 43.5 respectively.

Furthermore, we observe that the different dialects are not handled equally well by the models, even when we use localized models.[15] For example, Indian English are consistently worse than other varieties (cf. WER over 13), while perhaps unsurprisingly the best-handled English variety is the one from the United States.

The comparison between the "general" `en-US` model and the different localized models reveals interesting divergencies. When transcribing English from New Zealand, Nigeria, Scotland, South Africa, and Kenya, the localized models perform better than the US one. For Australian, Irish, and Philippine English, the differences between the two models are minor.

We also observe differences between the handling of Arabic varieties, with Algerian (DZA) proving particularly challenging. For both Bengali and Kiswahili, the two localized models produce exactly the same transcriptions, but the quality is consistent for two Bengali varieties around WER 27, with the Indian one being slightly better. On the other hand, the ASR quality of the two Swahili varieties exhibits the largest difference, with the Kenyan dataset receiving a 22% lower WER (cf. 37.9 and 49.1) than the Tanzanian one. Since the choice of the model does not affect the downstream WER for Bengali and Kiswahili, we will only use one for each language in subsequent experiments.

ASR Systems Unfairness We use the framework described in Section §3 to quantify the unfairness of the models. We will limit our discussion here on English, where we have more than one model to compare, but we provide extensive results in Appendix C. The two models we compare for English are the "general" one (obtained by using the en-US Google model) and a hypothetical "localized" one which assumes a pipeline involving dialect recognition (or a user-defined preference) that then selects the ASR model corresponding to the user's dialect.

Our results, summarily displayed in Table 4, are that the "general" English model gets an unfairness score of 0.0245, while the "localized" one obtains a lower unfairness score of 0.0221. This means that the "localized" ASR model is not only slightly better in terms of average transcription quality, but also leads to a more fair distribution of its benefits among the underlying users. Further analysis of the within-group and across-group unfairness scores of each region shows that both models achieve their lowest within-group unfairness scores for the US dialect (around 0.0014); this means that both models are more equitable for US English speakers than for speakers of other dialects. In contrast, the models have twice as high unfairness scores for South Indian English speakers (with unfairness scores around 0.0030). This means that not all South Indian English speakers receive consistent benefits (high quality transcriptions) by the systems.

Taking a look at the unfairness scores for the other languages, we observe that (a) all other models not only have worse ASR quality

[14]We will use the term "localized systems" as ones that are advertised as such to perform better on a specific language variety. We however do not have access to the actual training data to confirm this.

[15]We reiterate that because the dataset is *parallel* across the language varieties, these results are directly comparable.

Model	English Variety											
	AUS	IND-S	IND-N	IRE	NZL	NGA	PHI	SCO	ZAF	KEN	USA-SE	Avg
en-US	10.94	14.25	13.96	9.43	14.62	12.67	12.67	11.21	11.55	13.68	8.97	12.17
en-VAR	10.41	17.27	17.34	9.77	8.69	11.10	12.28	9.30	8.73	11.03		11.35

Model	Bengali Variety		
	BGD	IND	Avg
bn-BD	28.40	26.73	27.56
bn-IN	28.47	26.73	27.6

Model	Kiswahili Variety		
	KEN	TZA	Avg
sw-KE	37.89	49.10	43.49
sw-TZ	37.89	49.10	43.49

Model	Arabic Variety							
	DZA	BHR	EGY	JOR	MAR	SAU	TUN	Avg
ar-VARIETY	38.41	36.29	35.87	36.05	35.82	36.50	35.61	36.36

Table 3: Development WER (lower is better) on each of the language varieties, using different localized speech recognition models.

ASR Model	Unfairness score ↓	Avg. WER ↓
en-US	0.025	12.17
en-VAR	0.022	11.35
ar-XX	0.077	36.36
bn-XX	0.036	27.6
sw-XX	0.163	43.49

Table 4: A hypothetical English ASR model using localized ASR models is not only better in terms of average quality (WER) but also slightly more equitable than the "general" en-US model. The ASR systems for other languages are generally more unfair than the English ones.

but are also more unfair over their respective populations. Bengali receives an unfairness score of 0.0355 while Arabic receives scores around 0.08. Kiswahili, however, beyond being by far the worst in terms of WER, is also the most unfair system with double the unfairness score (0.1626) of the second most unfair language system. This is unsurprising, considering how wide the performance (WER) gap is between Kenyan and Tanzanian Swahili.

Sensitive Feature Analysis The metadata associated with each annotator in SD-QA allow us to perform analyses across sensitive features like gender or age. We provide a breakdown of WER across these two features for all varieties and ASR models in Tables 14 (age) and 13 (gender) in Appendix B. The tables also provide information on the support for each average score to allow for better interpretation of the results. We leave a more sophisticated analysis incorporating statistical significance tests for future work.

Studying the effect of the speaker's gender, we do not find large differences between the average WER for most varieties, but nevertheless we can make some interesting observations. First, we note that Indian English behaves differently depending on whether the speakers are from south or north India. In particular, female speakers from the south receive slightly higher (worse) WER than their male counterparts (c.f. WER of 14.4 and 12.9). For speakers from the Indian north the situation is reversed, with female speakers receiving almost half of the WER of males (c.g. 8.5 and 15.8). For both Scottish and US English, the utterances of female speakers seem to be easier to transcribe than male ones, with a difference of about 3 WER points in both cases. While in Bangladesh Bengali we do not observe significant differences between the average WER for female and male speakers, in Indian Bengali the difference is more than 5 WER points, with male WER being better (lower).

4.2 QA Tasks Results

In this section, we investigate the effect of ASR errors on the downstream QA tasks, also discussing the performance across different language varieties. We first present results on the development set, on which we base our discussion, before providing results on the test sets.

Passage Selection Task Table 5 lists the obtained F_1-SCORE over the SD-QA development set, using both the original questions ("Gold" in Table 5) which simulate a perfect ASR system, as well as the noisy transcriptions from the different ASR models.

In almost all cases, using the noisy ASR output question has detrimental effect to the task. The reduction in F_1-SCORE varies from just 0.3

Transc.	English Variety AUS	IND-S	IND-N	IRE	NZL	NGA	PHI	ZAF	KEN	SCO	USA-SE	Avg.
Gold	53.4											
en-US	51.1	50.9	50.2	51.9	52.4	49.9	50.3	51.6	48.7	51.5	52.8	51
en-VAR	52.8	48.8	48.8	52.2	54.2	51.8	50.9	52.9	51.3	52.6		51.7

Transc.	Bengali Variety BGD	IND	Avg.
Gold	57.9		
bn-XX	56.6	57.6	57.1

Transc.	Kiswahili Variety KEN	TZA	Avg.
Gold	68.8		
sw-XX	57.5	55.8	56.7

Transc.	Arabic Variety DZA	BHR	EGY	JOR	MAR	SAU	TUN	Avg.
Gold	65.0							
ar-VAR	65.2	64	63	63.6	65.6	63.8	64.5	64.2

Table 5: Baseline passage selection results (F_1-SCORE, higher is better) on the SD-QA development set.

Transc.	English Variety AUS	IND-S	IND-N	IRE	NZL	NGA	PHI	ZAF	KEN	SCO	USA-SE	Avg.
Gold	37.3											
en-US	33.5	32.3	32.8	34.0	34.9	32.5	33.5	32.4	32.3	33.8	34.6	33.3
en-VAR	35.7	31	30	34.0	34	33.3	30.9	35.1	33.3	34.8		33.3

Transc.	Bengali Variety BGD	IND	Avg.
Gold	47.7		
bn-XX	47.9	48.8	48.4

Transc.	Kiswahili Variety KEN	TZA	Avg.
Gold	56.8		
sw-XX	43.8	41.9	42.9

Transc.	Arabic Variety DZA	BHR	EGY	JOR	MAR	SAU	TUN	Avg.
Gold	51.3							
ara-VAR	51.3	46.2	44.8	45.6	47.6	46.3	46.7	46.9

Table 6: Baseline results (F_1-SCORE) on the minimal answer task on the SD-QA development set.

(in Indian Bengali) to 13 percentage points (in Tanzanian Swahili). The effect is generally less pronounced in Bengali (reduction of -0.8 points on average) and Arabic (-0.7 points), while it is clearly significant in Swahili (-12.1 points on average).

Interestingly, there are three cases where the “noisy" ASR transcripts lead to slightly better performance that the gold transcripts. These are New Zealand English (NZL) using the localized ASR model and Algerian (DZA) and Morrocan (MAR) Arabic.

Minimal Answer Task Table 6 presents the F_1-SCORE on the minimal answer task. As before, in most cases using the noisy ASR transcriptions lead to QA performance deterioration.

In English, we find no difference between using the transcriptions of the general or the localized ASR systems (both have a macro-average F_1-SCORE of 33.3), which contrasts with our findings for the passage selection task where the better transcriptions from localized ASR system's lead to slightly better downstream performance.

Notably, the downstream performance for both Bengali dialects is improved when the input is the output of the ASR system, with a stronger effect for Indian Bengali. We plan on studying this interesting result in future work. On the other side of the spectrum, performance in Swahili is significantly impacted, with an average reduction of almost 14 percentage points.

QA Systems Unfairness We use the framework described in Section §2 to quantify the unfairness of the QA systems with respect to their underlying populations. The results are listed in Table 7. The differences between languages are less pronounced. We observe the lowest unfairness score for Bengali speakers (~ 0.04), while for

Language	Passage Selection		Minimal Answer	
	Unfairness score ↓	Avg. F_1-SCORE ↑	Unfairness score ↓	Avg. F_1-SCORE ↑
en-VAR	0.082	51.0	0.078	33.3
ar-XX	0.076	64.2	0.076	46.9
bn-XX	0.043	57.1	0.047	48.4
sw-XX	0.089	56.7	0.090	42.9

Table 7: QA systems exhibit different levels of unfairness across languages, being more fair for Bengali speakers and less fair for Swahili or English speakers.

Language	Passage Sel.	Minimal Answer
ara	-0.29	-0.25
ben	-0.94	-0.94
eng	-0.92	-0.81
swa	-1.00	-1.00

Table 8: Speech transcription quality correlates with downstream QA accuracy (Spearman's rank correlation coefficient between WER and F_1-SCORE).

English and Swahili speakers the unfairness score is more than double the Bengali one, with unfairness scores over 0.08.

The downstream effect of ASR noise Perhaps unsurprisingly, we find that, *within each language*, the quality of the ASR transcription correlates with downstream QA performance, for both tasks. We calculate the Spearmans rank correlation coefficient for each language for both tasks, and present results in Table 8, also visualizing them in Figure 3.[16] For English, we use the localized en-VAR models.

We note that Arabic vernaculars, unlike the English, Bengali, and Swahili varieties, seem to exhibit a different behavior, as evidenced by the lower correlation coefficients (c.f. -0.25 to < -0.8 for other languages in the minimal answer task). We leave an analysis of this behavior for future work, but we believe it can be attributed to the fact that the original questions are in Modern Standard Arabic, which may bias the speakers of different vernaculars to a varying extent.

Test Set Results To facilitate future comparisons against both ASR and QA models, we also report results with our baseline pipeline approach. For English, we use the localized ASR models, as the analysis on the development set shows they are better than using the "general" US English model. Table 9 shows the ASR system's quality on the test set, Table 10 presents the results for all dialects on the passage selection task, while the results on the minimal answer task are listed in Table 11.

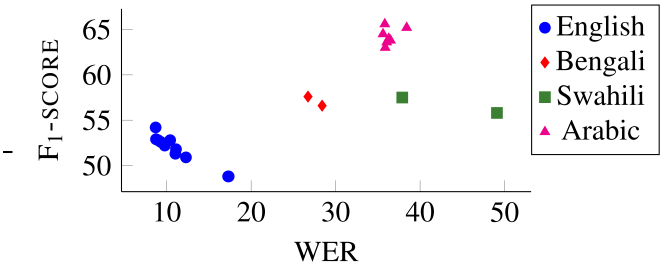
As in the development set, the noisy transcriptions lead to worse downstream performance compared to using the gold questions. Unlike the development set results though, we note that this hold for *all* languages and dialects for both tasks, even for e.g. Bengali (where the noisy transcriptions actually lead to slightly higher F_1-SCORE for the minimal answer task). In addition, the F_1-SCORE differences between the gold and noisy settings are generally larger than those we observed on the development set.

We refrain from performing any additional analysis on the test set, and suggest that any future analysis be conducted on the development set, so that all test set results reflect an estimation of real-world performance on *unseen* data.

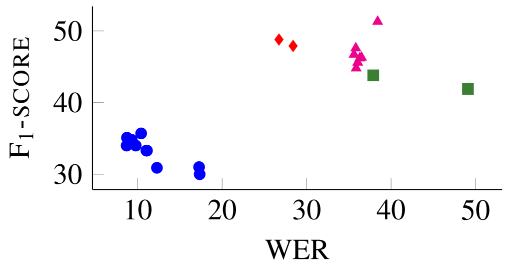
5 Related Work

Benchmark Datasets There exist a number of benchmark question answering datasets. SQuAD (Rajpurkar et al., 2016) provides a passage and a question which has an answer placed in the passage. In SQuAD 2.0 (Rajpurkar et al., 2018) additional unanswerable questions were introduced. However, in SQuAD the annotators first read the passage and create the questions based on it. A more realistic setting is to instead focus on "random" questions which might be asked to search engines without reading any passage. The Natural Questions (NQ) dataset (Kwiatkowski et al., 2019) attempts to address this gap, consisting of anonymized Google queries. Other datasets focus on a trivia (TriviaQA (Joshi et al., 2017)) or conversational setting (CoQA (Reddy, 1989) and QuAC (Choi et al., 2018)).

[16]The correlation coefficients are negative because for WER a lower score is better, while for F_1-SCORE a higher score is better.



(a) Passage Selection task



(b) Minimal Answer task

Figure 3: The downstream accuracy on the two QA task (F_1-SCORE) is generally negatively correlated to the question transcription quality (WER), for each language's dialects. This observation does not hold when comparing results across languages.

Model	AUS	IND-S	IND-N	IRE	NZL	NGA	PHI	SCO	ZAF	KEN	USA-SE	Avg
	English Variety											
en-US	10.43	11.65	10.13	9.25	8.79	15.35	8.32	11.35	8.57	12.81	5.25	10.17
en-VAR	8.54	16.48	13.29	8.91	7.27	11.61	8.95	8.14	7.56	10.24		9.66

Model	BGD	IND	Avg
	Bengali Variety		
bn-BD	30.87	32.30	31.59
bn-IN	30.83	32.30	31.56

Model	KEN	TZA	Avg
	Kiswahili Variety		
sw-KE	40.05	60.58	50.31
sw-TZ	40.05	60.58	50.31

Model	DZA	BHR	EGY	JOR	MAR	SAU	TUN	Avg
	Arabic Variety							
ar-VARIETY	37.84	37.93	40.40	38.55	39.09	38.24	38.70	38.68

Table 9: Test WER on each of the language varieties, using different localized speech recognition models.

Generally, conversational and dialog based QA datasets introduce new challenges as the questions are in free form and highly contextual. Notably, all the aforementioned datasets are only in English.

Multilingual QA Beyond monolingual QA models, cross-lingual QA systems aim to leverage resources from one language to answer questions originally asked in a different language. Recently released, the TyDi-QA dataset, upon which we build, contains question-passage-answer pairs in 11 typologically diverse languages. XOR-QA (Asai et al., 2020) explores the direction of open domain QA systems by introducing 3 cross lingual tasks. Asai et al. (2020) also build their cross lingual dataset on top of TyDi-QA questions, where questions originally unanswerable are associated with useful resources like English translations, related English wikipedia articles and any answers found are then translated to the original question language. Other notable benchmark cross-lingual QA datasets include MLQA (Lewis et al., 2019), MKQA (Longpre et al., 2020) and XQuAD (Artetxe et al., 2019). MLQA (Lewis et al., 2019) also explores cross lingual alignment among 7 language instead of training on monolingual large dataset, providing translations of the original English questions. MKQA (Longpre et al., 2020) provides the most diverse multilingual QA dataset comprised of 26 languages, with the data being translations of English NQ questions. Last, XQuAD (Artetxe et al., 2019) is another translated benchmark, with the questions being translations of the English ones from the SQuAd dataset. We opted for recording questions from the TyDi-QA dataset, in order to work with realistic questions for each language and avoid the effect of translationese.[17]

Speech QA A number of recent studies are done to bridge the gap between text based QA system and speech data. In Spoken SQuAD (Lee et al., 2018b), the authors propose a new task where the question is in textual form but the related reading comprehension is given in speech form. So transcription is performed on the speech data and the output can be in either text form or audio time span. Even using state-of-the-art speech transcription model, the authors show severe deterioration in performance. The

[17] See discussion on "Why not translate?" by Clark et al. (2020).

Transc.	English Variety											Avg
	AUS	IND-S	IND-N	IRE	NZL	NGA	PHI	ZAF	KEN	SCO	USA-SE	
Gold	58.7											
en-US	56.4	54.1	56.1	56	55.1	54.1	57.2	56.3	54.6	57	57.2	55.8
en-VAR	57.1	52	53.9	57.1	57.2	53.7	54.5	56.5	55.9	56.8		55.6

Transc.	Bengali Variety		Avg
	BGD	IND	
Gold	61.9		
bn-XX	60.3	59.7	60

Transc.	Kiswahili Variety		Avg
	KEN	TZA	
Gold	61.9		
sw-XX	43.8	41.4	42.6

Transc.	Arabic Variety							Avg
	DZA	BHR	EGY	JOR	MAR	SAU	TUN	
Gold	79.9							
ar-VAR	77.9	77	76.3	77.7	77.5	76.6	77.3	77.2

Table 10: Baseline passage selection results (F_1-SCORE, higher is better) on the SD-QA test set.

Transc.	English Variety											Avg
	AUS	IND-S	IND-N	IRE	NZL	NGA	PHI	ZAF	KEN	SCO	USA-SE	
Gold	37											
en-US	33.5	31.8	33.7	33.6	33.1	31.7	35.7	35	33.4	34	35.7	33.7
en-VAR	34.5	30.8	32	34.0	35.3	31.7	33.2	34	33	33.9		33.5

Transc.	Bengali Variety		Avg
	BGD	IND	
Gold	47.9		
bn-XX	47.2	45.8	46.5

Transc.	Kiswahili Variety		Avg
	KEN	TZA	
Gold	49.8		
sw-XX	31.4	29.3	30.35

Transc.	Arabic Variety							Avg
	DZA	BHR	EGY	JOR	MAR	SAU	TUN	
Gold	66.4							
ara-VAR	62.3	62.1	61.1	62.3	62.3	61.7	62.4	62.0

Table 11: Baseline results on the minimal answer task (F_1-SCORE, higher is better) on the SD-QA test set.

authors further propose subword unit sequence embedding based mitigation strategies. This work was further extended to the ODSQA dataset (Lee et al., 2018a), where the question is also given in speech form.

The most relevant work on the intersection of speech and QA is the recent NoiseQA (Ravichander et al., 2021) study, which explores the effect of noise in QA systems introduced by 3 main input interfaces: keyboard input, machine translation and speech input. Our task of assessing speech input error is conceptually similar to their speech input assessment. Their experiments show that the absence of punctuation in ASR outputs results in performance degradation by 5.1%. In addition, voice variation, accent and speaker's acoustic conditions as well as choice of ASR unit also play an important role. This study also shows that transcription of naturally spoken question results in errors like question type shift, ungrammatical or meaningless questions, corrupted named entities and dropped delimiters. Ravichander et al. (2021) make a number of recommendations including assessing the source of error, context-driven evaluation and community priorities while designing robust QA systems.

We believe that our work is a necessary expansion and complement of such studies. We augment Ravichander et al. (2021) by providing real audio data in more than one language instead of synthetic text-to-speech data generated from simulation experiments. Furthermore, we go beyond English by providing data in three more languages and several varieties/locales.

6 Conclusion

We present SD-QA, a new benchmark for the evaluation of QA systems in real-world settings, mimicking the scenario of a user querying a QA system through a speech interface. SD-QA is the largest spoken only multilingual and multi-dialect QA dataset to date, with coverage of four languages and twenty-two dialects.

We provide baseline performance and fairness results on a pipeline that uses publicly available ASR models to transcribe spoken queries before passing them to a multilingual QA system. We showcase the QA systems' lack of robustness to noise in the question. We also discuss the *fairness* of both speech recognition and QA models with regards to underlying user characteristics, and show that a users' dialect can significantly impact the utility they may receive from the system.

Future areas of improvement for this work include expanding SD-QA to cover more languages and dialects, and additional analysis on attributes about the data and users. Ideally, we would like to discern which parameters most influence the performance of downstream language systems. We also plan to investigate the prospect of an end-to-end spoken QA system in an attempt to bridge the gap between the speech modality of the query and the textual modality of the currently available knowledge bases.

Acknowledgments

This work is generously supported by NSF Award 2040926. The dataset creation was supported by Google through an Award for Inclusion Research. We also want to thank Jacob Eisenstein, Manaal Faruqui, and Jon Clark for helpful discussions on question answering and data collection. The authors are grateful to Kathleen Siminyu for her help with collecting Kiswahili and Kenyan English speech samples, to Sylwia Tur and Moana Wilkinson from Appen for help with the rest of the data collection and quality assurance process, and to all the annotators who participated in the creation of SD-QA.

References

Abdel-Rahman Abu-Melhim. 1991. Code-switching and linguistic accommodation in arabic. In *Perspectives on Arabic Linguistics III: Papers from the Third Annual Symposium on Arabic Linguistics*, volume 80, pages 231–250. John Benjamins Publishing.

Chris Alberti, Kenton Lee, and Michael Collins. 2019. A BERT baseline for the natural questions.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations.

Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2020. XOR QA: Cross-lingual open-retrieval question answering. arXiv:2010.11856.

Anton Batliner, Ralf Kompe, A Kießling, E Nöth, and H Niemann. 1995. Can you tell apart spontaneous and read speech if you just look at prosody? In *Speech Recognition and Coding*, pages 321–324. Springer.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models? In *Proc. AAAI*.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Elodie Gauthier, David Blachon, Laurent Besacier, Guy-Noel Kouarata, Martine Adda-Decker, Annie Rialland, Gilles Adda, and Grégoire Bachman. 2016. LIG-AIKUMA: a Mobile App to Collect Parallel Speech for Under-Resourced Language Studies.

Nizar Y Habash. 2010. Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *Proceedings of the 55th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Mimi H Klaiman. 1987. Bengali. *The World's Major Languages, Croon Helm, London and Sydney*, pages 490–513.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee. 2018a. ODSQA: Open-domain spoken question answering dataset. *2018 IEEE Spoken Language Technology Workshop (SLT)*.

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018b. Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension. *Interspeech 2018*.

Patrick Lewis, Barlas Ouz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. MLQA: Evaluating cross-lingual extractive question answering.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. MKQA: A linguistically diverse benchmark for multilingual open domain question answering.

Colin P Masica. 1993. *The indo-aryan languages*. Cambridge University Press.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text.

Abhilasha Ravichander, Siddharth Dalmia, Maria Ryskina, Florian Metze, Eduard Hovy, and Alan W Black. 2021. NoiseQA: Challenge set evaluation for user-centric question answering.

John Rawls. 1999. *A Theory of Justice*. Harvard University Press.

Raj Reddy. 1989. Speech research at Carnegie mellon. In *Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod, Massachusetts, October 15-18, 1989*.

Candide Simard, Sarah M Dopierala, and E Marie Thaut. 2020. Introducing the sylheti language and its speakers, and the soas sylheti project. *Language Documentation and Description*, 18:1–22.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248.

Hanne-Ruth Thompson. 2020. *Bengali: A comprehensive grammar*. Routledge.

A Dataset Details

We provide detailed statistics on the number of examples in the development and test sets, as well as on the speakers for each language/dialect in Table 12.

B ASR Results Breakdown

B.1 By Gender

Results in Table 13.

B.2 By Age

Results in Table 14.

C Detailed Unfairness Results

We present a complete breakdown of the unfairness calculations for each language and dialect on the ASR task. For each dialect, we report two numbers. First, the within-group unfairness score, which can be interpreted as answering the question “how fair/consistent is the ASR model for the speakers of this dialect?". Second, we report the between-group unfairness score, which provides information on the average benefit a subgroup/dialect receives relative to all other dialects; a negative value means that this dialect is treated unfairly with respect to the rest.

Table 15 presents results in English, Table 18 presents results in Arabic, Table 16 presents results in Swahili, and Table 17 presents results in Bengali.

Language	Dialect	Development Set Examples	Development Set Speakers (M,F)	Test Set Examples	Test Set Speakers (M,F)
Arabic	Algeria (DZA)	708	4 (2, 2)	1380	7 (5, 2)
	Bahrain (BHR)	708	3 (2, 1)	1380	7 (5, 2)
	Egypt (EGY)	708	3 (2, 1)	1380	6 (4, 2)
	Jordan (JOR)	708	3 (2, 2)	1380	7 (4, 3)
	Morocco (MAR)	708	4 (3, 1)	1380	7 (4, 3)
	Saudi Arabia (SAU)	708	3 (2, 1)	1380	7 (5, 2)
	Tunisia (TUN)	708	3 (2, 1)	1380	8 (4, 4)
Bengali	Bangladesh (BGD)	427	14 (8, 6)	328	11 (5, 6)
	India (IND)	427	4 (3, 1)	326	2 (1, 1)
English	Australia (AUS)	1000	4 (2, 2)	1031	5 (1, 4)
	India-South (IND-S)	1000	5 (4, 1)	1031	4 (4, 0)
	India-North (IND-N)	1000	4 (3, 1)	1031	4 (–, –)
	Ireland (IRL)	1000	4 (2, 2)	1031	5 (0, 5)
	Kenya (KEN)	972	4 (4, 0)	984	4 (2, 2)
	New Zealand (NZL)	1000	6 (3, 3)	1031	5 (3, 2)
	Nigeria (NGA)	1000	5 (3, 2)	1031	6 (2, 4)
	Philippines (PHI)	1000	4 (2, 2)	1031	4 (1, 3)
	Scotland (SCO)	1000	4 (2, 2)	1031	5 (1, 4)
	South Africa (ZAF)	1000	5 (3, 2)	1031	6 (1, 5)
	USA-Southeast (USA-SE)	1000	4 (2, 2)	1031	5 (2, 3)
Kiswahili	Kenya (KEN)	2299	15 (7, 8)	2184	8 (7, 1)
	Tanzania (TZN)	2150	13 (6, 7)	1675	6 (3, 3)

Table 12: Data and annotator statistics for SD-QA.

Model	Gender	English Variety AUS	IND-S	IND-N	IRE	NZL	NGA	PHI	ZAF	KEN	SC0	USA-SE
en-US	Female	10.92 (446)	14.36 (867)	8.55 (247)	9.1 (517)	10.28 (372)	15.16 (455)	12.4 (535)	12.3 (512)	–	9.46 (434)	6.92 (457)
	Male	10.96 (526)	12.93 (105)	15.82 (725)	10.34 (455)	8.9 (600)	14.15 (517)	12.89 (437)	10.7 (460)	13.61 (912)	12.58 (538)	10.83 (515)
en-VAR	Female	9.65 (446)	17.03 (867)	11.66 (247)	8.88 (517)	8.86 (372)	11.49 (455)	12.59 (535)	9.23 (512)	–	7.52 (434)	
	Male	11.06 (526)	18.6 (105)	19.24 (725)	10.73 (455)	8.58 (600)	10.72 (517)	11.93 (437)	8.18 (460)	11.13 (912)	10.69 (538)	

Model	app:details	Bengali Variety BGD	IND
bn-BD	Female	28.46 (180)	31.58 (63)
	Male	28.38 (247)	25.86 (364)
bn-IN	Female	28.46 (180)	31.58 (63)
	Male	28.49 (247)	25.86 (364)

Model	Gender	Kiswahili Variety KEN	TZA
sw-KE	Female	32.72 (1990)	40.1 (1137)
	Male	30.9 (165)	42.91 (895)
sw-TZ	Female	32.72 (1990)	40.1 (1137)
	Male	30.9 (165)	42.91 (895)

Model	Gender	Arabic Variety DZA	BHR	EGY	JOR	MAR	SAU	TUN
ar-VAR	Female	39.68 (168)	35.01 (270)	38.64 (168)	33.81 (270)	34.67 (672)	36.48 (540)	34.2 (270)
	Male	37.07 (540)	35.9 (438)	36.64 (540)	36.22 (438)	41.7 (36)	33.37 (168)	35.28 (438)

Table 13: Development WER (example count grouped by speaker gender) for each of the language varieties, using dialect specific and general speech recognition.

Transc.	Age	AUS	IND-S	IND-N	IRE	NZL	NGA	PHI	ZAF	KEN	SCO	USA-SE
		English Variety										
en-US	18-30	9.53 (186)	16.18 (452)	18.18 (540)	7.46 (190)	9.92 (517)	13.49 (592)	12.13 (702)	11.86 (80)	13.68 (972)	–	7.96 (455)
	31-45	12.35 (529)	12.54 (520)	8.58 (432)	10.26 (782)	8.91 (440)	16.3 (380)	–	11.04 (892)	–	11.18 (972)	9.88 (517)
	46-59	8.99 (257)	–	–	–	7.92 (15)	–	13.95 (270)	–	–	–	–
en-VAR	18-30	10.04 (186)	19.12 (452)	19.71 (540)	7.74 (190)	8.49 (517)	10.82 (592)	12.19 (702)	11.86	11.04 (972)	–	
	31-45	11.76 (529)	15.57 (520)	14.21 (432)	10.28 (782)	8.91 (440)	11.45 (380)	–	8.43 (892)	–	9.27 (972)	
	46-59	7.82 (257)	–	–	–	8.91 (15)	–	12.53 (270)	–	–	–	

Transc.	Age	Bengali Variety BGD	IND
bn-BD	18-30	28.71 (397)	25.7 (270)
	31-45	24.78 (30)	27.92 (151)
	46-59	–	39.53 (6)
bn-IN	18-30	28.78 (397)	25.7 (270)
	31-45	24.78 (30)	27.92 (151)
	46-59	–	39.53 (6)

Trans.	Age	Kiswahili Variety KEN	TZA
sw-KE	18-30	–	–
	31-45	27.44 (830)	32.94 (614)
	46-59		
sw-TZ	18-30	–	–
	31-45	27.44 (830)	32.94 (614)
	46-59	–	–

Model	Age	DZA	BHR	EGY	JOR	MAR	SAU	TUN
		Arabic Variety						
ara-VAR	18-30	39.68 (168)	35.01 (270)	–	35.32 (708)	35.06 (708)	36.48 (540)	32.91 (438)
	31-45	37.07 (540)	32.34 (168)	36.33 (270)	–	–	33.37 (168)	37.81 (270)
	46-59	–	37.92 (270)	37.63 (438)	–	–	–	–

Table 14: Development WER (example count by speaker age-group) for each of the language varieties, using dialect specific and general speech recognition.

Model	Unfairness Component	AUS	IND-S	IND-N	IRE	NZL	NGA	PHI	SCO	ZAF	KEN	USA-SE	Total Unfairness
		English Variety											
en-US	within group	0.00186	0.00293	0.00277	0.00172	0.00165	0.00282	0.00239	0.00197	0.00224	0.00243	0.00148	0.02456
	between group	0.00118	-0.00284	-0.00235	0.00242	0.00279	-0.00326	-0.00076	0.00112	0.000218	-0.00169	0.00347	
en-VAR	within group	0.00175	0.00308	0.00291	0.00159	0.00151	0.00217	0.00207	0.00155	0.00159	0.00179	0.00146	0.02212
	between group	0.00117	-0.00633	-0.00609	0.00185	0.00301	-0.00001	-0.00119	0.00236	0.00276	0.00037	0.00277	

Table 15: Unfairness Components for ASR in English.

Model	Unfairness Component	Kiswahili Variety KEN	TZA	Total Unfairness
sw-KE	within group	0.0744	0.084179	0.162634
	between group	0.047031	-0.042977	
sw-TZ	within group	0.0744	0.084179	0.162634
	between group	0.047031	-0.042977	

Table 16: Unfairness Components for ASR in Kiswahili.

Model	Unfairness Component	Bengali Variety BGD	IND	Total Unfairness
bn-BD	within group	0.018628	0.016795	0.0355
	between group	-0.006184	0.006262	
bn-IN	within group	0.018718	0.016803	0.035601
	between group	-0.006305	0.006385	

Table 17: Unfairness Components for ASR in Bengali.

Model	Unfairness Component	DZA	BHR	EGY	JOR	MAR	SAU	TUN	Total Unfairness
		Arabic Variety							
ar-VAR	within group	0.011998	0.010899	0.011314	0.010912	0.010653	0.010494	0.010604	0.077006
	between group	-0.003639	0.000272	-0.002942	0.001526	0.001984	-0.00011	0.00304	

Table 18: Unfairness Components for ASR in Arabic.